sensors

MDPI

Article

Multiscale Representation of Observation Error Statistics in Data Assimilation

Vincent Chabot [1,†], Maëlle Nodet [1,2] and Arthur Vidard [1,*]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2] Université Paris-Saclay, UVSQ, CNRS, Laboratoire de Mathématiques de Versailles, 78000 Versailles, France
* Correspondence: arthur.vidard@inria.fr
† Current address: Météo-France, Toulouse 31000, France.

Received: 20 December 2019; Accepted: 4 March 2020; Published: 6 March 2020

Abstract: Accounting for realistic observation errors is a known bottleneck in data assimilation, because dealing with error correlations is complex. Following a previous study on this subject, we propose to use multiscale modelling, more precisely wavelet transform, to address this question. This study aims to investigate the problem further by addressing two issues arising in real-life data assimilation: how to deal with partially missing data (e.g., concealed by an obstacle between the sensor and the observed system), and how to solve convergence issues associated with complex observation error covariance matrices? Two adjustments relying on wavelets modelling are proposed to deal with those, and offer significant improvements. The first one consists of adjusting the variance coefficients in the frequency domain to account for masked information. The second one consists of a gradual assimilation of frequencies. Both of these fully rely on the multiscale properties associated with wavelet covariance modelling. Numerical results on twin experiments show that multiscale modelling is a promising tool to account for correlations in observation errors in realistic applications.

Keywords: data assimilation; observation errors; error correlation; multiscale analysis; wavelets; error covariance matrices

1. Introduction

Numerical weather prediction requires the determination of the initial state of the system in order to produce forecasts. Retrieving an optimal initial condition requires the use of so-called data assimilation methods that combine information from observations, model equations and their respective error statistics.

Since the late 1970s satellites have been a dominant source of information. Errors associated with such data are highly correlated in space and between different frequency channels, which can be detrimental if they are not accounted for, even approximately [1,2]. Due to the size of the observation vectors, building and handling corresponding error covariance matrices is not feasible in practice. Consequently most data assimilation systems assume that observations are uncorrelated with each other. This either induces severe misspecification of error statistics, or necessitates the use of only a fraction of the available observations to ensure this assumption to be valid [3]. Considering the high cost of remote sensing observation, this situation should be avoided. For this reason the representation of correlated observation errors has very recently become a significant topic of research and several routes are being explored. First research was directed to accounting of inter-channel error correlation. Due to the modest size of the resulting covariance matrix, the main problems lie in the poor quality of the error estimations and on the detrimental effect it has on the conditioning of the minimisation problem (see [4,5] and bibliography). For spatial correlation, a practical approach is to use a covariance matrix that is block diagonal. This is manageable when observations can be grouped into small enough

batches which are uncorrelated with each other [6]. For a more general spatial distribution [7] proposes to represent convolutions with the covariance matrix by a diffusion equation discretized using a finite element approach.

On the other hand, when observation are dense in space, some (multiscale) transformations can be applied to the data in order to perform efficient subsampling of the observations [3]. Such transformations can also be used to permit a cheap but good approximation of said error statistics representation [2]. The latter approach, which is the main topic of this paper suffers from two main difficulties. First, dealing with partially missing data in one set of observation is not straightforward and requires a special treatment of observation error statistics in the frequency domain. Second, as mentioned above, considering spatially correlated observation errors can severely damage the convergence properties of the assimilation methods. In this paper, after a short introduction to the context of general data assimilation (Section 2.1) and wavelet representation of the observation errors (Section 2.2), we present its actual implementation in the main data assimilation techniques (Sections 2.3 and 2.4) and discuss in detail the aforementioned difficulties Section 2.5. Proposed solutions are then implemented on a simple case mimicking a laboratory experiment (presented in Section 2.6), and their performance is discussed in Section 3.

2. Materials and Methods

2.1. General Formulation of Data Assimilation

Let $\mathcal{M}$ be a model describing of the dynamics in a given system, represented by its state vector $\mathbf{x}$. For example, $\mathbf{x}$ might be a vector of temperatures over a grid (discretized area of interest).

$$\frac{\partial \mathbf{x}}{\partial t}(t) = \mathcal{M}(\mathbf{x}(t)), \quad \mathbf{x}|_{t=0} = \mathbf{x}_0 \tag{1}$$

where $\mathbf{x}_0$ is the initial value of the state vector.

Data assimilation aims at providing an analysis $\mathbf{x}^a$ which will be used to compute optimal forecasts of the system's evolution.

Such an analysis is produced using various sources of information about the system: observations (measurements), previous forecasts, past or a priori information, statistics on the data and/or model errors, and so on.

In this paper, we assume that these ingredients are available:

- the numerical model $\mathcal{M}$,
- a priori information about $\mathbf{x}_0$, denoted $\mathbf{x}_0^b$ and called *background state vector*,
- partial and imperfect observations of the system, denoted $\mathbf{y}$ and called *observation vector*,
- the observation operator $\mathcal{H}$, mapping the state space into the observation space,
- statistical modelling of the background and observation errors (assumed unbiased), by means of their covariance matrices $\mathbf{B}$ and $\mathbf{R}$.

Data assimilation provides the theoretical framework to produce an optimal (under some restrictive hypotheses) analysis $\mathbf{x}^a$ using all the aforementioned ingredients. In this work, we will focus on how to make the most of the observation error statistics information and we will not consider the background error information. Regarding the observation information, typically, most approaches can be formulated as providing the best (in some sense) vector in order to minimize the following quantity, measuring the misfit to the available information:

$$\|\mathcal{H}(\mathbf{x}) - \mathbf{y}\|_{\mathbf{R}}^2 \tag{2}$$

where the notation $\|\mathbf{z}\|_{\mathbf{K}}^2$ stands for the Mahalanobis distance; namely $\|\mathbf{z}\|_{\mathbf{K}}^2 = \mathbf{z}^T\mathbf{K}^{-1}\mathbf{z}$. Some information about algorithms and methods will be given in following paragraphs. For an extensive description we refer the reader to the recent book [8].

2.2. Spatial Error Covariance Modelling Using Wavelets

Being able to accurately describe the covariances matrices $\mathbf{B}$ and $\mathbf{R}$ is a crucial issue in data assimilation, as they count as main ingredients in the numerical computation. The $\mathbf{B}$ matrix modelling has been largely investigated (see e.g., [9,10]). DA works actually using non diagonal $\mathbf{R}$ matrices are quite recent (e.g., [2,7,11]). Evidence shows that observation errors are indeed correlated [12] and that ignoring it can be detrimental [13,14].

In [2] the authors introduced a linear change of variable $\mathbf{A}$ for accounting for correlated observation errors, while still using a diagonal matrix in the algorithm core. For the sake of clarity we will summarize the approach in the next few lines. If we assume that the observation error ϵ is such that $\epsilon = \mathbf{y} - \mathbf{y}^t$, with $\epsilon \sim \mathcal{N}(0, \mathbf{R})$, $\mathbf{y}^t$ being the true vector (without any error) and $\mathcal{N}(0, \mathbf{R})$ designing the normal distribution of zero mean and covariance matrix $\mathbf{R}$. Then changing variables writes $\beta = \mathbf{A}\epsilon = \mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{y}^t$ and $\beta \sim \mathcal{N}(0, \mathbf{A}\mathbf{R}\mathbf{A}^T)$. Then we carefully choose $\mathbf{A}$ so that the transformed matrix is almost diagonal: $\mathbf{D}_\mathbf{A} = \mathrm{diag}(\mathbf{A}\mathbf{R}\mathbf{A}^T) \simeq \mathbf{A}\mathbf{R}\mathbf{A}^T$. Indeed, we then have the following property:

$$\begin{aligned}\|\mathbf{y} - \mathcal{H}(\mathbf{x})\|^2_\mathbf{R} &= (\mathbf{y} - \mathcal{H}(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x})) \\ &\simeq (\mathbf{y} - \mathcal{H}(\mathbf{x}))^T\mathbf{A}^T\mathbf{D}_\mathbf{A}^{-1}\mathbf{A}(\mathbf{y} - \mathcal{H}(\mathbf{x})) = \|\mathbf{A}\mathbf{y} - \mathbf{A}\mathcal{H}(\mathbf{x})\|^2_{\mathbf{D}_\mathbf{A}}\end{aligned}$$

After this change of variable, the covariance matrix that will be used in the data assimilation algorithm is therefore $\mathbf{D}_\mathbf{A}$, it is diagonal. At the same time, the covariance information still has some interesting features, if the change of variable $\mathbf{A}$ is carefully chosen.

As an illustration, Figure 1 presents the correlations of the central point with respect to its neighbors for diagonal covariance matrices using various changes of variables: none, change into wavelet space, change into Fourier space, change into curvelet space. This figure was produced using a diagonal correlation matrix $\mathbf{D}$, then applying the chosen change of variable to obtain $\mathbf{R} = \mathbf{A}\mathbf{D}\mathbf{A}^T$, then plotting the correlation described by $\mathbf{R}$. We can see in the figure that interesting correlations can be produced with an adequate change of variable. Indeed, all these changes of variables have the following fact in common: they perform a change of basis such that the new basis vectors have supports distributed over multiple neighboring points (contrary to the classical Euclidean basis vector, which are zero except in one point). This fact explains the fact that $\mathbf{R}$ is now non-diagonal.

Let us explain briefly the Fourier, wavelet and curvelet change of variables. For Fourier, the image is decomposed in the Fourier basis:

$$\mathbf{y} = \sum_j \langle \mathbf{y}, \varphi_j \rangle \varphi_j$$

where (φ_j) represents the Fourier basis (e.g., sinusoidal functions) and the index j describes the scale of the jth basis vector (think of j as a frequency). The change of variables consists of describing $\mathbf{y}$ by its coefficients y_j on the basis (φ_j): $y_j = \langle \mathbf{y}, \varphi_j \rangle$.

Similarly for the wavelets, the decomposition writes

$$\mathbf{y} = \sum_{j,k} \langle \mathbf{y}, \varphi_{j,k} \rangle \varphi_{j,k}$$

where $(\varphi_{j,k})$ represents the wavelet basis (e.g., Haar or Daubechies), where the index j describes the scale of the jth basis vector and k is its position in space (think of wavelets as localised Fourier functions). The change of variables ($\mathbf{A}$, denoted $\mathbf{W}$ for the wavelets) into wavelets space consists of describing $\mathbf{y}$ by its coefficients $y_{j,k}$ on the basis $(\varphi_{j,k})$: $y_{j,k} = \langle \mathbf{y}, \varphi_{j,k} \rangle$. In other words, $\mathbf{W}\mathbf{y}$ is the vector of coefficients $(y_{j,k})_{j,k}$.

This is also similar for the curvelets:

$$\mathbf{y} = \sum_{j,k,l} \langle \mathbf{y}, \varphi_{j,k,l} \rangle \varphi_{j,k,l}$$

where the index l describe the orientation of the basis vector.

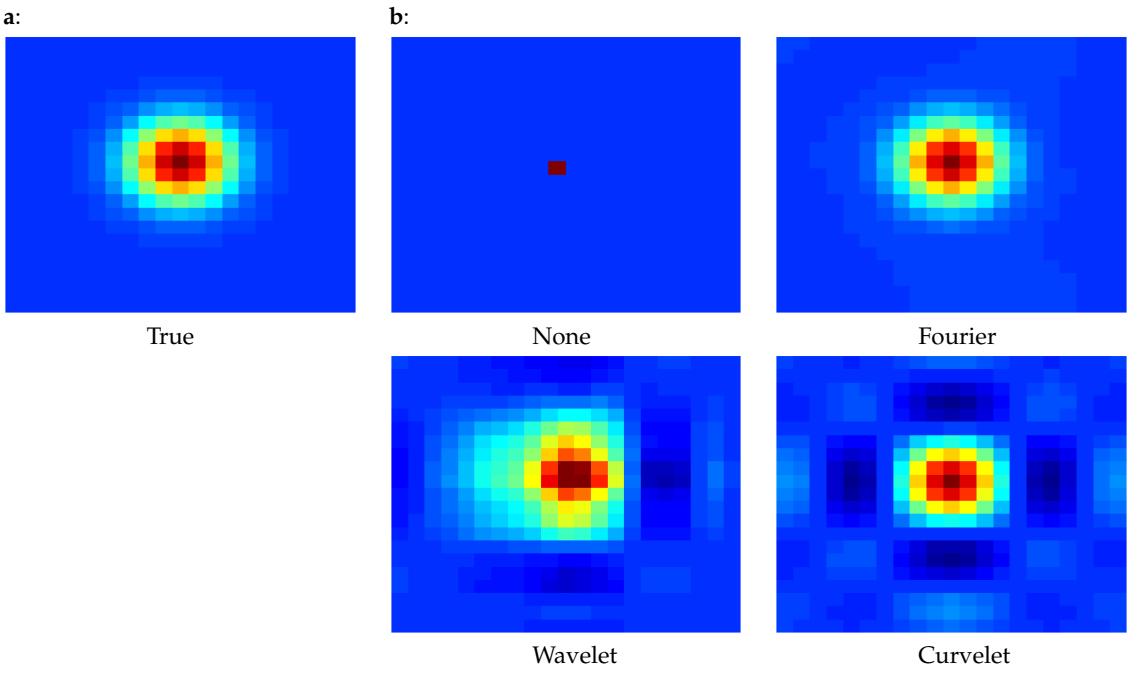

Figure 1. Correlation of the central observation point with respect to its neighbors. Dark red indicates values close to 1, blue is 0. (**a**) true correlation that we are trying to reproduce. (**b**) correlations obtained with the combination of a diagonal matrix and four different changes of variable: none (**top left**), change into Fourier space (**top right**), change into wavelet space (**bottom left**), change into curvelet space (**bottom right**).

Using these changes of variables then allows various observation error modelling:

- Fourier: when the errors change with the scale only
- Wavelets: when the errors change with the scale as well as the position (e.g., for a geostationary satellite whose incidence angle impacts the errors, so that the errors vary depending on the position in the picture)
- Curvelets: when the errors change with the scale, the position and the orientation (e.g., when errors are highly non linear and depend on the flow, so that they are more correlated in one direction than another).

In this work, our focus is with wavelet basis, which presents many advantages: there exists fast wavelet transform algorithms (as for Fourier), so the computational cost remains reasonable. Also, contrary to Fourier, wavelets are localised in space and allow error correlations that are inhomogeneous in space, which is more realistic for satellite data, as well as data with missing zones.

To be more specific about wavelet transform, let's assume the observation space is represented by a subset of $\mathbb{Z}$, where each number represents a given observation point location (in 1D). Wavelet decomposition consists of computing, at each given scale, a coarse approximation at that scale, and finer details. Both are decomposed on a multiscale basis and are therefore represented by their coefficients on the bases. Approximation and details coefficients are given by a convolution formula:

$$c^{j-1}[n] = \sum_{p\in\mathbb{Z}} h[p-2n]c^{j}[p]; \quad d^{j-1}[n] = \sum_{p\in\mathbb{Z}} g[p-2n]d^{j}[p]$$

where $c^j[n]$ represents the approximation coefficient at scale j at point $n \in \mathbb{Z}$, $d^j[n]$ represents the details coefficient at scale j at point $n \in \mathbb{Z}$, h and g are functions depending on the chosen wavelets basis, each of them being equal to zero outside of their support $[n_1; n_2]$. Moreover, wavelet g has k

vanishing moments. A wavelet is said to have a vanishing moment of order k if g is orthogonal to every polynomial of degree up to $k-1$. As an example a wavelet with 1 vanishing moment is represented by a filter g such that $\sum_n g[n] = 0$. This property is very important. Indeed, if the correlation is smooth enough (i-e. can be well approximated by polynomials of degree smaller than k), then details coefficients have a very small variance.

This can be extended in 2D (or more), where details coefficients at scale j will be separated into 3 components: vertical (d_v^j), horizontal(d_h^j) and diagonal(d_d^j). Bottom-left panel of Figure 2 shows the classical representation of coefficients on the wavelet space of a 2D signal. Finer details coefficient being stored in the three larger submatrices. The coarse approximation at finer scale is stored in the top-left submatrix and is itself decomposed into details and a coarser approximation. In this example the signal is decomposed into three scales.

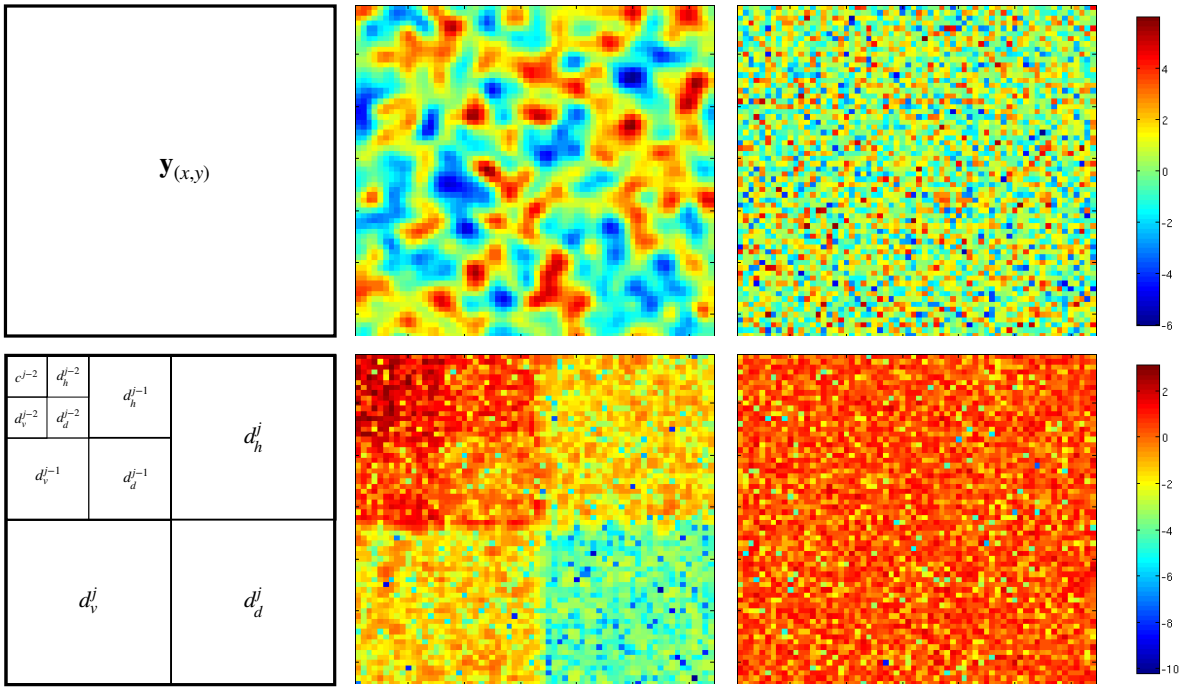

Figure 2. Top middle (resp right) panel shows an example of correlated (resp uncorrelated) noise. On bottom left, the scheme of the organisation of wavelet coefficient with three scales. On bottom middle (resp right), the logarithm of the absolute value in wavelet space of the correlated (resp uncorrelated) noise. We can see that approximation coefficient (c^{j-2}) are significantly larger than small scale details coefficients (d_*^j).

The top row of Figure 2 shows examples of both correlated (middle panel) and uncorrelated (right panel) noise. The bottom row shows their respective coefficient (in log-scale) in wavelet space using the representation depicted above. While uncorrelated noise affect all scales indiscriminately, the effect of correlated noise is significantly different from one scale to another (up to a factor 100 in this example). One can observe that approximation coefficients are very large compare to small scale details coefficients. This means that correlated noise (or smooth noise) affect more approximation coefficients than small scale details coefficients. This is due to the "vanishing moment" property of the wavelet g. Additionnaly the effect of a correlated noise resemble a (different) uncorrelated noise on each scale, meaning the diagonal approximation of the error covariance matrix will be a good one, as long as the sub-diagonals corresponding to each scales are different. This is represented in Figure 3 that shows the variances (log-scale) in the wavelet space of both correlated and uncorrelated noise from Figure 2.

In the next paragraphs we will describe hos this transformation can be used in the two classical frameworks of Data Assimilation: variational methods and filtering methods.

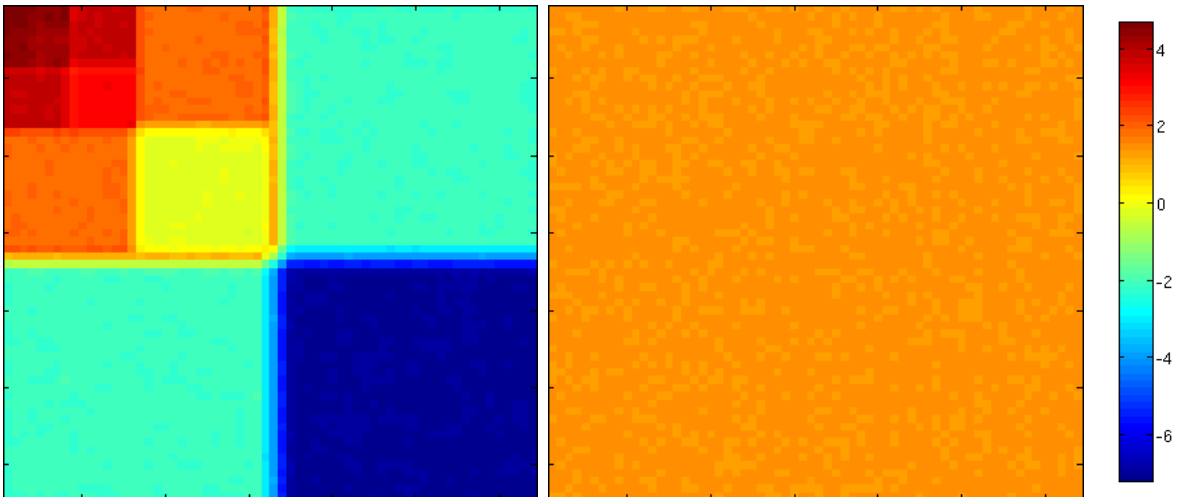

Figure 3. Logarithm of the variances of a correlated (**left**) and an uncorrelated (**right**) noise in the wavelet space. One can see that approximation coefficients (for a correlated noise) have a very small variance while approximation coefficients have a large variance. When no correlation exists in the noise, all the coefficients have the same variance (which is one in this example).

2.3. Implementation in Variational Assimilation

In the framework of variational assimilation, the analysis is set to be the minimizer of the following cost function J, which diagnoses the misfit between the observations and a priori information and their model equivalent, as in (2):

$$J(\mathbf{x}_0) = J^b(\mathbf{x}_0) + J^o(\mathbf{x}_0), \quad \text{where} \quad J^b = \|\mathbf{x}_0 - \mathbf{x}_0^b\|^2_{\mathbf{B}}, \quad J^o = \|\mathcal{H}(\mathbf{x}(\mathbf{x}_0)) - \mathbf{y}\|^2_{\mathbf{R}}$$

where $\mathbf{x}(\mathbf{x}_0)$ is the solution of equation (1), when the initial state is $\mathbf{x}_0$. In practice $\mathbf{y}$ stores time-distributed observations, so that it can be written as

$$J^o(\mathbf{x}_0) = \sum_{i \text{ obs. time}} \|\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{y}_i\|^2_{\mathbf{R}_i}$$

where $\mathcal{H}_i$ is the observation operator at time i, $\mathbf{y}_i$ is the observation vector at this time, and $\mathbf{R}_i$ is the observation error covariance matrix.

Using the wavelets change of variables $\mathbf{A} = \mathbf{W}$, we choose a diagonal matrix $\mathbf{D}_w$ (possibly varying with the observation time i, but we omit the index for the sake of simplicity), and we set

$$J^o(\mathbf{x}_0) = \sum_{i \text{ obs. time}} \|\mathbf{W}\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{W}\mathbf{y}_i\|^2_{\mathbf{D}_w} \tag{3}$$

so that the observation error covariance matrix that is actually defined is:

$$\mathbf{R}^{-1} = \mathbf{W}^T\mathbf{D}_w^{-1}\mathbf{W} \tag{4}$$

Meanwhile, the algorithm steps are:

1. Compute the model trajectory $\mathbf{x}(\mathbf{x}_0)$ and deduce the misfits to observation $\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{y}_i$ for all i
2. Apply the change of variable (wavelet decomposition) $\mathbf{W}\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{W}\mathbf{y}_i$
3. Compute the contribution to the gradient for all i: $\nabla J^o = \mathbf{H}^T\mathbf{W}^T\mathbf{D}_w^{-1}(\mathbf{W}\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{W}\mathbf{y}_i)$
4. Descent and update following the minimization process

In this algorithm, we can see that there is no need to form nor inverse $\mathbf{R}$, the optimization module only sees the diagonal covariance matrix $\mathbf{D}_w$, so that the minimization can be approached with classical

methods like conjugate gradient or quasi Newton. Therefore, the only modification consists of coding the wavelets change of variable and its adjoint. As wavelet transforms are usually implemented using optimized and efficient libraries, the added cost is reasonable [2].

2.4. Implementation in Kalman Filtering

In this Section we explain the practical implementation of accounting for correlated observation errors in the Kalman filtering framework. We will briefly recall the main equations of the filters and then explain the adequate alterations to include observation error covariance modelling. We will use the standard notations and algorithms of data assimilation [8,15].

2.4.1. Extended Kalman Filter

Using standard notation, the analysis step of the extended Kalman filter writes as follows:

$$\begin{aligned} \mathbf{x}^a &= \mathbf{x}^f + \mathbf{K}(\mathbf{y}^o - \mathcal{H}(\mathbf{x}^f)) \\ \mathbf{K} &= \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1} \end{aligned} \tag{5}$$

where $\mathbf{x}^a$ and $\mathbf{x}^f$ are the analysis and forecast state vectors respectively, $\mathbf{y}^o$ is the observation vector, $\mathcal{H}$ is the (possibly nonlinear) observation operator, $\mathbf{H}$ is its tangent linearized version, $\mathbf{K}$ is the Kalman gain matrix, $\mathbf{P}^f$ is the forecast error covariance matrix and $\mathbf{R}$ is the observation error covariance matrix.

To account for a non diagonal $\mathbf{R}$ while keeping the algorithm easy to implement, let us assume that we define $\mathbf{R}$ as previously by (4), with $\mathbf{D}_w$ a diagonal matrix whose dimension is d, the number of wavelet coefficients (equal to the observation space dimension p). We recall that the wavelet transform $\mathbf{W}$ is orthonormal, we have

$$\mathbf{W}^T\mathbf{W} = \mathbf{I}_d, \qquad \mathbf{W}^{-1} = \mathbf{W}^T$$

where $\mathbf{I}_d$ is the identity matrix in dimension $d = p$. Then we can write the Kalman gain matrix as:

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\,\mathbf{W}^T\left(\mathbf{W}\mathbf{H}\mathbf{P}^f\mathbf{H}^T\mathbf{W}^T + \mathbf{D}_w\right)^{-1}\mathbf{W}$$

In this equation, we can see that the algorithm complexity is preserved, up to a change of variable:

- the required matrix inversion $(\mathbf{W}\mathbf{H}\mathbf{P}^f\mathbf{H}^T\mathbf{W}^T + \mathbf{D}_w)^{-1}$ can be expected to be of the same complexity as $(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}$,
- two changes of variables, and their inverse, are required, one on the matrix $\mathbf{H}\mathbf{P}^f\mathbf{H}^T$, and one in the end for the matrix $(\mathbf{W}\mathbf{H}\mathbf{P}^f\mathbf{H}^T\mathbf{W}^T + \mathbf{D}_w)^{-1}$ to get the final Kalman gain.

As wavelet transforms are usually implemented using optimized and efficient libraries, we can also expect the added cost to be affordable. In particular efficient parallel libraries exist [16] and it is well suited for GPU computing.

2.4.2. Stochastic Ensemble Kalman Filter

Let us recall the main ingredients of analysis step of the stochastic ensemble Kalman filter, as can be found in [8] (pp. 158–160). The number m stands for the number of members in the ensemble.

- A set of m perturbed observations is generated to account for the observation error:

$$\text{for } i = 1..m, \text{ compute} \quad \mathbf{y}_i^o = \mathbf{y}^o + \mathbf{u}_i, \quad \text{with } \mathbf{u}_i \sim \mathcal{N}(0, \mathbf{R})$$

- The $\mathbf{Y}_f$ matrix is computed. First we compute

$$\overline{\mathbf{u}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{u}_i$$

then the ith column of $\mathbf{Y}_f$ is given by:

$$\left[\mathbf{Y}_f\right]_i = \frac{\mathcal{H}(\mathbf{x}_i^f) - \mathbf{u}_i - (\mathcal{H}(\overline{\mathbf{x}^f}) - \overline{\mathbf{u}})}{\sqrt{m-1}}, \text{ for } i = 1..m$$

- The Kalman gain matrix is computed:

$$\mathbf{K} = \mathbf{X}^f \mathbf{Y}_f^T \left(\mathbf{Y}_f \mathbf{Y}_f^T\right)^{-1}$$

- The analysis members are computed:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}\left[\mathbf{y}_i^o - \mathcal{H}(\mathbf{x}_i^f)\right], \text{ for } i = 1..m$$

This algorithm can accommodated the accounting for observation errors correlations using this simple modification of the very first step (perturbation of the observations), which is the only occurrence of the $\mathbf{R}$ matrix:

$$\text{for } i = 1..m, \quad \begin{cases} \beta_i \sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{u}_i = \mathbf{W}^T \mathbf{D}_w^{1/2} \beta_i \\ \mathbf{y}_i^o = \mathbf{y}^o + \mathbf{u}_i \end{cases}$$

As we can see, this is quite easy to implement, as $\mathbf{D}_w$ is diagonal, its square root is easily obtained. $\mathbf{W}^T$ is the inverse of the wavelet transform. This operation has to be performed m times, possibly in parallel, so that the added cost should be negligible.

2.4.3. Deterministic Ensemble Kalman Filter

As previously, let us recall the main ingredients of the deterministic Kalman filter (from [8] p. 162 and following), and see how it can be adapted to account for a change of variable in wavelet space for observation error covariance modelling. Let us first take a look at the analysis phase of the filter:

- Contrary to the stochastic Kalman filter, the $\mathbf{Y}_f$ observation anomalies matrix is not perturbed:

$$\left[\mathbf{Y}_f\right]_i = \frac{\mathcal{H}(\mathbf{x}_i^f) - \overline{\mathbf{y}^f}}{\sqrt{m-1}} \quad \text{for } i = 1..m, \qquad \text{with} \quad \overline{\mathbf{y}^f} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{H}(\mathbf{x}_i^f)$$

- The analysis is given by

$$\mathbf{x}^a = \overline{\mathbf{x}^f} + \mathbf{K}\left[\mathbf{y}^o - \mathcal{H}(\overline{\mathbf{x}^f})\right], \text{ for } i = 1..m$$

- The Kalman gain matrix writes

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T \left(\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R}\right)^{-1}$$

- The analysis anomalies are given by

$$\mathbf{w}^a = \mathbf{Y}_f^T (\mathbf{Y}_f \mathbf{Y}_f^T + \mathbf{R})^{-1} \delta, \quad \text{with } \delta = \mathbf{y}^o - \mathcal{H}(\overline{\mathbf{x}^f})$$

which can be rewritten using an adapted version of the Sherman-Morrison-Woodbury formula:

$$\mathbf{w}^a = (\mathbf{I}_m + \mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f)^{-1} \mathbf{Y}_f^T \mathbf{R}^{-1} \delta = \mathbf{T}\, \mathbf{Y}_f^T \mathbf{R}^{-1} \delta, \text{ with } \mathbf{T} = (\mathbf{I}_m + \mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f)^{-1}$$

- And finally, here is the generation of the posterior ensemble:

$$\mathbf{x}_i^a = \overline{\mathbf{x}^f} + \mathbf{X}^f \left(\mathbf{w}^a + \sqrt{m-1} \left[\mathbf{T}^{\frac{1}{2}} U \right]_i \right)$$

where $\mathbf{X}^f$ is a matrix whose columns are the normalized forecast anomalies and $\mathbf{U}$ is an arbitrary orthogonal matrix:

$$[\mathbf{X}^f]_i = \frac{\mathbf{x}_i^f - \overline{\mathbf{x}^f}}{\sqrt{m-1}}, \text{ for } i = 1..m$$

The required modifications to include the change of variable $\mathbf{W}$ are twofold:

1. First, we change variable in the expression $\mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f$:

$$\mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f = \mathbf{Y}_f^T \mathbf{W}^T \mathbf{D}_w^{-1} \mathbf{W} \mathbf{Y}_f = (\mathbf{W}\mathbf{Y}_f)^T \mathbf{D}_w^{-1} (\mathbf{W}\mathbf{Y}_f)$$

Here we can see that the matrix inversion still occurs with a diagonal matrix, as before, and we just have to apply the wavelet transform on $\mathbf{Y}_f$, which is a matrix with m columns, where m is small.

2. Second, we change variable in the expression $\mathbf{Y}_f^T \mathbf{R}^{-1} \delta$:

$$\mathbf{Y}_f^T \mathbf{R}^{-1} \delta = (\mathbf{W}\mathbf{Y}_f)^T \mathbf{D}_w^{-1} (\mathbf{W}\delta)$$

Here the change of variable is done only once on δ the inovation vector. It has been done previously for $\mathbf{Y}_f$.

Notice that the matrix inversion $\mathbf{T} = (\mathbf{I}_m + \mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f)^{-1}$, as well as computing its square-root $\mathbf{T}^{\frac{1}{2}}$, takes place in ensemble subspace, of dimension m, and is therefore efficient even if the change of variable impacted $\mathbf{Y}_f^T \mathbf{R}^{-1} \mathbf{Y}_f$.

2.5. Toward Realistic Applications

This approach works well with idealistic academic test cases. To go toward realistic applications, several issues need to be sorted out. In this section we address two of them. The first one is quite general and requires the ability to deal with incomplete observations, where part of the signal is missing, either due to sensor failure or external perturbation/obstruction. The second one is more specific to variational data assimilation, where the conditioning of the minimisation, and hence its efficiency, can be severely affected by complex correlation structure in the observation error covariance matrix. It is likely to also affect the Kalman Filter, in particular the matrix inversion in the observation space it requires (e.g., in Equation (5)), but it is yet to be demonstrated.

2.5.1. Accounting for Missing Observations

When dealing with remote sensing, reasons for missing observation are numerous, ranging from a glitch in the acquisition process to an obstacle blocking temporarily one part of the view. This may be quite detrimental to our proposed approach since it violates the multi-scale decomposition hypotheses. However, contrary to Fourier, wavelets (and many, if not all, x-lets) have local support that may be exploited to handle this issue. Please note that the same kind of issue can arise in case of complex geometry. For instance if one observes sea surface temperatures, land is present in the observation, while not being part of the signal. Somehow it can be treated as missing value.

One possibility would be to use inpainting techniques to fill in the missing values. However, this would make the associated error very difficult to describe. Indeed, it would require the estimation of the errors associated with introducing 'new' data in the missing zones, which is likely to be of different nature than that of original observations.

The idea is therefore to adapt the $\mathbf{R}$ matrix to the available data. Without any change of variable, the adaptation would be straightforward, as we would just have to apply a projection operator π to both the data and the $\mathbf{R}$ matrix:

$$\mathbf{y}^o - \mathcal{H}(\mathbf{x}^f) \rightarrow \pi(\mathbf{y}^o_\pi - \mathcal{H}(\mathbf{x}^f)); \quad \mathbf{R} \rightarrow \pi \mathbf{R} \pi^T$$

where the projector π maps the full observation space into the subset of the observed points, and $\mathbf{y}^o_\pi$ represents the full observation vector (with 0 where there is no available data).

When using a change of variable into wavelet space, it is a bit more tricky to perform, as a given observation point is used to compute many wavelet coefficients. Vice-versa a given wavelet coefficient is based on several image observation points. As a consequence, if some observation points are missing and others are available, it may result in "partially observed" wavelet coefficients, as schematized in Figure 4.

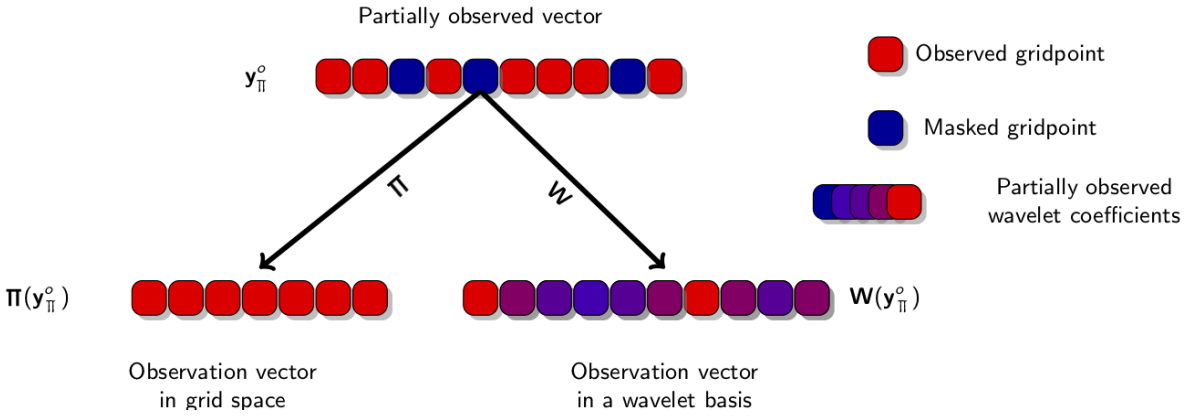

Figure 4. Schematic representation of the impact of non-observed pixels on the observation vector for both "pixel-grid" space and wavelet space. In pixel-grid space, missing pixels simply disappear. In wavelet space, missing pixels lead to partially observed wavelets coefficients.

Our choice is to still take into account these coefficients (and not discard them, because it would result in discarding too much information, as a single missing observation point affects numerous wavelet coefficients), but to carefully tune the diagonal coefficient of the diagonal matrix $\mathbf{D}_w$.

Missing observation points have two opposite effects:

- A missing observation point does not have signal nor any error, so we could expect the error variance of the impacted wavelet coefficients to decrease.
- A missing observation point leads to more discontinuities in the error signal as the error is 0 where a observation point is missing. This will increase significantly the fine scale coefficients, which is quite unfortunate since a good property of the wavelet decomposition was to have very small coefficient on finer scale (see end of Section 2.2).

To account for both effects, we propose an heuristic to adjust the variance σ^2_π (in other words, the coefficients of the diagonal matrix $\mathbf{D}_w$) corresponding to coefficients whose support is partially masked as follows:

$$\sigma^2_\pi = \left((1-\beta)\sigma^2 + \beta \sigma^2_a \right) I^2$$

where:

- σ^2 is the original error variance (e.g. given by the data provider);
- $\beta \in [0,1]$ is multiplied by the variance of the wavelet coefficient without any correlation σ_a: it accounts to inflating the variance due to missing information (loss of the error averaging effect);
- $I \in [0,1]$ stands for information content, and models the deflation effect. It takes into account the impact of missing observation points on the considered wavelet coefficient.

We now explain how β and I can be tuned. For the sake of simplicity, let us assume that our observation lives in a one dimensional space.

Computation of I.

The deflation percentage I for each coefficient is computed using also a wavelet transform, where h and g are replaced by constant functions with the same support:

$$h^0[n] = \frac{1}{n_2 - n_1 + 1}; \quad g^0[n] = \frac{1}{n_2 - n_1 + 1}; \quad \forall n \in [n_1; n_2]$$

Those functions extract the percentage of observation point present on the wavelet support. We proceed as follows. First we set the mask corresponding to the missing observation, it is an observation vector m equals to 1 where the observation point is observed and equal to zero where the observation point is missing. The wavelet transform of the mask aims to keep track of the impact of any missing observation point on any given wavelet coefficient. The percentage I is computed for each coefficient $c^j[n]$ and $d^j[n]$ by induction:

- At the finest scale j_{max}:

$$\begin{cases} I(c^{j_{max}}[n]) &= \sum_{p\in\mathbb{Z}} h^0[p-2n]m[p] \\ I(d^{j_{max}}[n]) &= \sum_{p\in\mathbb{Z}} g^0[p-2n]m[p] \end{cases}$$

- At the other scales:

$$\begin{cases} I(c^{j-1}[n]) &= \sum_{p\in\mathbb{Z}} h^0[p-2n]I(c^j[p]) \\ I(d^{j-1}[n]) &= \sum_{p\in\mathbb{Z}} g^0[p-2n]I(c^j[p]) \end{cases}$$

Computation of β.

Let us now explain how to compute the inflation coefficient β. As explained previously, as g has k vanishing moments, small scale coefficients have small variances. However, when using masked signal, one loses this property. In other world, missing data damages the smoothness of the signal (and of the noise), which in turn damages the efficiency of wavelet representation. The coefficient β reflect the loss of the first vanishing moment: in the following formula we can see that β is zero if the first vanishing moment is preserved, and non-zero if not, in order to inflate the variance of small scales. For the finest scale, β is given by:

$$\begin{cases} c_m^{j_{max}}[n] &= \sum_{p\in\mathbb{Z}} \frac{|h[p-2n]|}{\sum_{q\in\mathbb{Z}}|h[q-2n]|} m[p] \\ \beta_m^{j_{max}}[n] &= \left|\sum_{p\in\mathbb{Z}} \frac{g[p-2n]\,m[p]}{\sum_{q\in\mathbb{Z}}|g[q-2n]|\,m[q]}\right| \end{cases}$$

Indeed, $\sum_{p\in\mathbb{Z}} g[p-2n]m[p] = 0$ means that wavelet still has a 0-th order null moment, even with missing coefficients, and in that case $\beta = 0$.

Coarser scales coefficients are computed by induction, as:

$$\begin{cases} c_m^{j-1}[n] &= \sum_{p\in\mathbb{Z}} \frac{|h[p-2n]|}{\sum_{q\in\mathbb{Z}}|h[q-2n]|} c_m^j[p] \\ \beta_m^{j-1}[n] &= \left|\sum_{p\in\mathbb{Z}} \frac{g[p-2n]\,m[p]}{\sum_{q\in\mathbb{Z}}|g[q-2n]|\,m[q]} c_m^j[p]\right| \end{cases}$$

Finally, the variance model is modified as follows for every detail coefficient whose data is partially missing:

$$\sigma_\pi^2(d^j[n]) = \left(\sigma^2 + \beta_m^j[n]\sigma_a^2\right)\, I(d^j[n])^2$$

For approximation coefficient, only the deflation factor is used:

$$\sigma_\pi^2(c^j[n]) = \sigma^2\, I(d^j[n])^2$$

Indeed, when the error is correlated, the variance of the approximation coefficient σ^2 is much greater than σ_a^2. This is the case on Figure 3 where $\sigma^2 \sim 100$ while $\sigma_a^2 = 1$. Moreover, h can be seen as a local smoothing operator ($\sum_n h[n] = 1$) and therefore correlated errors do not compensate themselves. Consequently, their is no need for inflation. Inversely, for finer details, in our example $\sigma^2 \sim 10^{-2}, 10^{-4}$ so β has a significant impact on those scales.

These modifications give therefore a new diagonal matrix $\mathbf{D}_w$ which takes into account the occurrence of missing information. Section 3 will present numerical results.

2.5.2. Gradual Assimilation of the Smallest Scales

As will be shown in the numerical results Section 3 below, another issue can occur with real data: convergence issues due to the nature of observation errors. Indeed, what our experiments highlight is that our test-case behaves well when the represented error correlation are Gaussian and homogeneous in space. For correlated Gaussian errors whose correlations are inhomogeneous in space, convergence issues occur to the point that it destroys the advantage of using wavelets: they do worse than the classical diagonal matrix without correlation. Please note that in a general case, even accounting for homogeneous noise may degrade the conditioning of the minimization [4]. Wavelet transform does not change the conditioning of the problem, but its multi-scale nature can be of help to circumvent this problem.

Numerical investigation of the results shows that some sort of aliasing occurs for small wavelet scales. Indeed, smallest scales are the least affected by the correlated noise, so they are not well constrained by the assimilation and they tend to cause a divergence when large scales are not well known either, which is at the beginning of the assimilation iteration process. Removing the smaller scales altogether is not a suitable solution, as they contain valuable information we still want to use. The proposed solution is therefore to first assimilate the data without the small scales and then add smaller scales gradually. Please note that this is not a scale selection method per se, as all scales will eventually be included. It can be related to the quasi-static approach [17] that gradually include observations over time.

Description of the gradual scale assimilation method.

Let us rewrite the observation cost function given by Equation (3):

$$\begin{aligned} J^o(\mathbf{x}_0) &= \sum_{i \text{ obs. time}} \|\mathbf{W}\mathcal{H}_i(\mathbf{x}(\mathbf{x}_0)) - \mathbf{W}\mathbf{y}_i\|^2_{\mathbf{D}_w} \\ &= \sum_{i \text{ obs. time}} \sum_{s \text{ scale}} \sum_k \frac{|d^s_{\mathbf{y}_i}[k] - d^s_{\mathcal{H}_i(\mathbf{x})}[k]|^2}{\sigma^2_{s,k}} \end{aligned}$$

where $d^s_{\mathbf{y}_i}[k]$, for $k \in \mathbb{Z}$, (resp. $d^s_{\mathcal{H}_i(\mathbf{x})}[k]$) represent the wavelet coefficients at scale s of the signal $\mathbf{y}_i$ (resp. $\mathcal{H}_i(\mathbf{x})$) and the $\sigma^2_{s,k}$ are the associated error variances (corresponding to the diagonal coefficients of the matrix $\mathbf{D}_w$).

Let us denote by $J^o_{s,i}$ the total cost corresponding to the scale s and observation time i:

$$J^o_{s,i} = \sum_k \frac{|d^s_{\mathbf{y}_i}[k] - d^s_{\mathcal{H}_i(\mathbf{x})}[k]|^2}{\sigma^2_{s,k}}$$

We then decide that the information at a given scale is usable only if the cost remains small, e.g. smaller than a given threshold τ_s, we define the thresholded cost J^o_{s,i,τ_s} by:

$$J^o_{s,i,\tau_s} = \begin{cases} J^o_{s,i} & \text{if } J^o_{s,i} \leq \tau_s \\ \tau_s & \text{otherwise} \end{cases}$$

The new observation cost function is then:

$$J^o_{\tau_s}(\mathbf{x}_0) = \sum_{i \text{ obs. time}} \sum_{s \text{ scale}} J^o_{s,i,\tau_s}$$

As mentioned before, the same issue could arise when using Kalman Filter type techniques during the matrix inversion needed when computing the gain matrix. Similar approaches based on iterative and multi-resolution could be used to sort this out.

2.6. Experimental Framework

Numerical experiments have been performed to study and illustrate the two issues that were previously highlighted: how to account for covariances with missing observations, and how to improve the algorithm convergence while still accounting for smaller scale information. This paragraph describes the numerical setup which has been used.

We wish to avoid adding difficulty to these already complex issues, therefore we chose a so-called *twin experiment* framework. In this approach, synthetic observations are created from a given state of the system (which we call the "true state", which will serve as reference) and then used in assimilation.

The experimental model represents the drift of a vortex on the experimental turntable CORIOLIS (Grenoble, France), which simulates atmospheric vortices in the atmosphere: the turning of the table provides an experimental environment which emulates the effect of the Coriolis force on a thin layer of water. A complete rotation of the tank takes 60 seconds, which corresponds to one Earth rotation.

2.6.1. Numerical Model

A numerical model represents the experiment, using the shallow-water equations on the water elevation $h(x, y, t)$ and the horizontal velocity of the fluid $\mathbf{w}(x, y, t) = (u(x, y, t), v(x, y, t))$, where u and v are the zonal and meridional components of the velocity. The time variable t is defined on an interval $[t_0, t_f]$, while the space variable (x, y) lives in Ω a rectangle in the plane $\mathbb{R}^2$. The equations write:

$$\begin{cases} \partial_t u - (f + \zeta)v + \partial_x B &= -ru + \kappa\Delta u \\ \partial_t v + (f + \zeta)u + \partial_y B &= -rv + \kappa\Delta v \\ \partial_t h + \partial_x(hu) + \partial_y(hv) &= 0. \end{cases}$$

The relative vorticity is denoted by $\zeta = \partial_x v - \partial_y u$ and the Bernoulli potential by $B = gh + \frac{u^2+v^2}{2}$, where g is the gravity constant. The Coriolis parameter on the β-plane is given by $f = f_0 + \beta y$, κ is the diffusion coefficient and r the bottom friction coefficient. The following numerical values were used for the experiments: $r = 9.10^{-7}$, $\kappa = 0$, $f_0 = 0.25$, $g = 9.81$ and $\beta = 0.0406$. The model is discretized using a finite differences scheme over a 128×128 grid and a 4^{th}-order Runge-Kutta scheme in time, with a time step of 2.5s. Please note that this means the model fields can be decomposed in up to 7 different scales using wavelet transform ($128 = 2^7$).

Additional equations represent the evolution of the tracer concentration (fluorescein):

$$\begin{cases} \partial_t q + \nabla q \cdot \mathbf{w} - \nu_T \Delta q = 0 \\ q(t_0) = q_0. \end{cases} \tag{6}$$

where q_0 is the initial concentration of the tracer (assumed to be known), $\nu_T = 10^{-5}$ is the tracer diffusion coefficient and $\mathbf{w} = (u, v)$ the fluid velocity computed above.

2.6.2. Synthetic Observations for Twin Experiments

In the twin experiment framework, observations are computed using the model. A known "true state" is used to produce a sequence of images which constitutes the observations. Therefore, the observation operator $\mathcal{H}$ is given by:

$$\mathcal{H}(\mathbf{x}_i) = q(t_i). \tag{7}$$

where $q(t_i)$ comes from (6).

Then assimilation experiments are performed starting from another system state, using synthetic observations. The results of the analysis can then be compared to the synthetic truth.

Unless otherwise stated, the assimilation period will be of 144 min, with one snapshot of passive tracer concentration every 6 min (24 snapshot in total). A selection of such snapshots is shown in Figure 5.

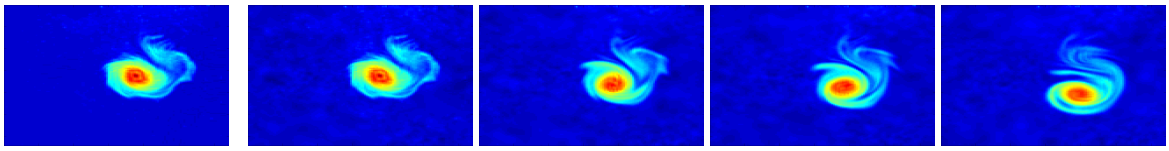

Figure 5. "True" initial concentration of the passive tracer (**1st left**) and noisy observations at initial time (**2nd**), after 90 min (**3rd**), 150 min (**4th**) and 270 min (**right**).

The observations are then obtained by adding an observation error $\mathbf{y} = \mathbf{y}^t + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbf{R})$ and $\mathbf{R}$ a suitably chosen matrix.

Our experiments will focus on three different formulations of the observation error covariance matrix. We will refer to "Pixels" the experiments for which there is no change of variable and the observation error covariance matrix is equal to $\mathbf{D} = \text{diag}(\mathbf{R})$. "Wavelet" will represent the experiments with the wavelet change of variable $\mathbf{W}$ and the observation error covariance matrix $\mathbf{D}_w = \text{diag}(\mathbf{WRW}^T)$. Finally, the last set of experiments will proceed as for the wavelets but will adjust the observation error covariance matrix according to the computations presented in Sections 2.5.1 and 2.5.2. The following table 1 summarises this up.

Table 1. Summary of the experiments description: name, change of variable, observation error covariance matrix.

Experiment Name	Change of Variable	Observation Error Covariance Matrix
Pixels	none (identity)	$\mathbf{D} = \text{diag}(\mathbf{R})$
Wavelet	$\mathbf{W}$	$\mathbf{D}_w = \text{diag}(\mathbf{WRW}^T)$
Wavelet tweaked	$\mathbf{W}$	$\mathbf{D}_w$ modified according to Section 2.5.1
Wavelet scale by scale	$\mathbf{W}$	$\mathbf{D}_w$ modified according to Section 2.5.2

3. Results

3.1. Accounting for Missing Observations

Figure 6 provides an example of image data with 10% missing observations. It represents three images from an temporal observation sequence, in which we simulated the presence of a passing cloud. This sequence has been generated using the experimental model presented above, and the masking cloud is advected at a regular pace.

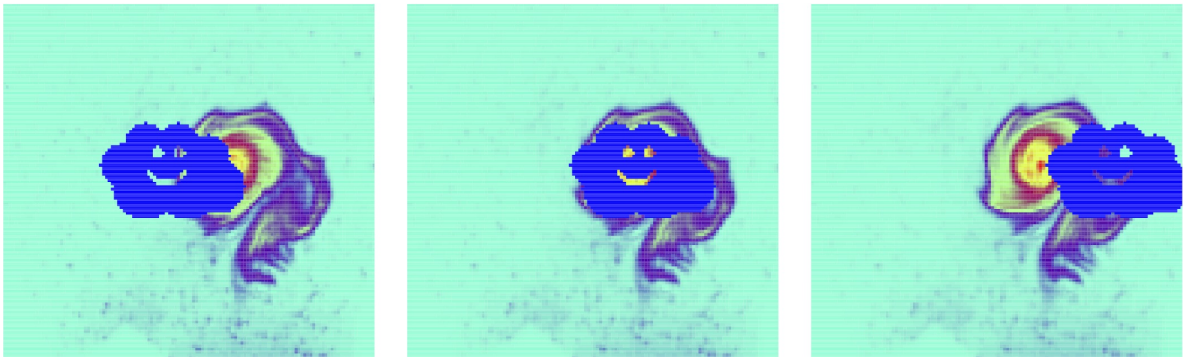

Figure 6. Example of an image sequence of the passive tracer concentration with missing observations. Left: first observation in the sequence, right: last observation.

This image sequence was then modified by a strong additive and spatially correlated homogeneous and isotropic noise (signal to noise ratio SNR = 14.8 dB). Then we performed many twin data assimilation experiments, while varying two parameters:

- the covariance error matrix: diagonal in observation space, diagonal in wavelet space (no adjustment), diagonal in wavelet space and modified according to Section 2.5.1 (see Table 1);
- the percentage of occulted signal: varying from 0 to 18% (with varying cloud sizes). For each experiment, the passing cloud has the same shape but different sizes.

For each experiment, we computed τ the ratio between the root mean square error for the analysis and the background:

$$\tau = \frac{\text{RMSE(analysis)}}{\text{RMSE(background)}} = \frac{\|(h_0^t, \mathbf{w}_0^t) - (h_0^a, \mathbf{w}_0^a)\|}{\|(h_0^t, \mathbf{w}_0^t) - (h_0^b, \mathbf{w}_0^b)\|}$$

where $(h_0^t, \mathbf{w}_0^t)$, $(h_0^a, \mathbf{w}_0^a)$ and $(h_0^b, \mathbf{w}_0^b)$ represent the true, analysed and background initial states of the experimental system. This ratio is close to zero when the analysis is much closer to the true state than the background (which represents the "no assimilation" state), and close to 1 when the analysis performs poorly. Figure 7 shows the resulting ratios for all the experiments. We can draw the conclusion that modifying the covariance matrix as proposed allows a considerable improvement from other methods, as it keeps the error below 20%, even for a widely occulted image sequence, despite the high noise level.

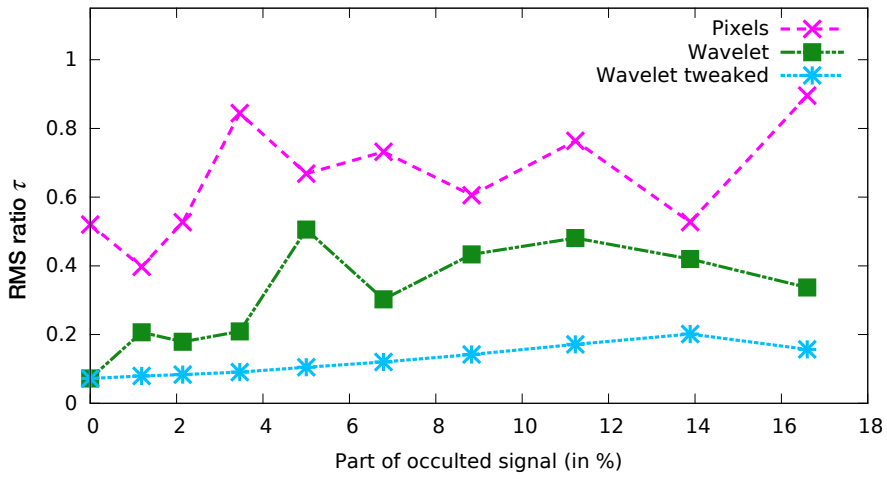

Figure 7. Ratio τ between the analysed state RMSE and the background RMSE for observations distorted by a strong spatially correlated additive noise, as a function of the percentage of missing observation points.

Figure 8 gives more details for the experiments with 9% occulted signal, as it represents (as a function of the spatial variable $x \in \Omega \subset \mathbb{R}^2$, see Section 2.6.1 for more details) the errors $v^t(x,0) - v^a(x,0)$, where $v^t(x,0)$ is the true longitudinal velocity at time 0 and $v^a(x,0)$ is the analysed longitudinal velocity at time 0. From this figure we can confirm that the modified wavelet covariance matrix does a much better job in approximating the true state.

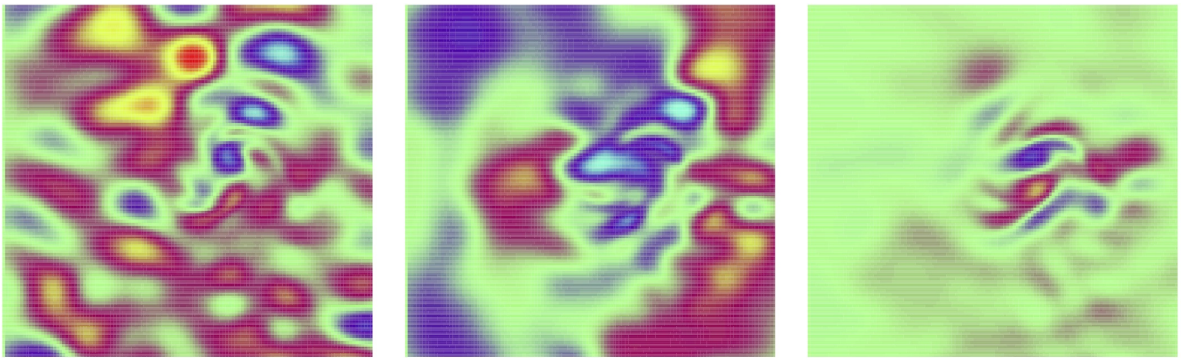

Figure 8. Error between the true velocity v and the analysed velocity after assimilation of an observation sequence with strong spatially correlated additive noise, with 9% missing observation points. The color scale ranges from -0.01 m.s^{-1} (**blue**) to 0.01 m.s^{-1} (**red**), which amounts to a third of the maximum velocity (ranging from -0.028 m.s^{-1} to 0.028 m.s^{-1}). Left: result for the pixel method, middle: wavelet without modification, right: improved wavelet method.

3.2. Gradual Assimilation of the Smallest Scales

Figure 9 illustrates the issue that we try to tackle using gradual assimilation. This figure presents the ratio r_k of the residual errors, as a function of the iteration number k:

$$r_k = \frac{\|(h_0^t, \mathbf{w}_0^t) - (h_0^k, \mathbf{w}_0^k)\|}{\|(h_0^t, \mathbf{w}_0^t) - (h_0^b, \mathbf{w}_0^b)\|}$$

As before $(h_0^t, \mathbf{w}_0^t)$ and $(h_0^b, \mathbf{w}_0^b)$ represent the true and background initial states of the experimental system. Index k represents the iteration number (loop index in the assimilation process) and $(h_0^k, \mathbf{w}_0^k)$ is the initial state vector computed by the assimilation system after k iterations. Both panels of Figure 9 shows the evolution of these ratio as a function of k for the "Pixels" and the "Wavelet" methods for covariance matrices modelling, as described in Table 1. The difference lies in the actual error that is added to the observations:

- on the left panel, the error is as previously described: $\mathbf{y} = \mathbf{y}^t + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbf{R})$, it is spatially correlated but the correlation is homogeneous in space ;
- on the right panel, we added an inhomogeneoulsy correlated error : $\epsilon = \mathbf{W}^T \mathbf{D}_w^{1/2} \beta$ with $\beta \sim \mathcal{N}(0, \mathrm{I})$.

As we can see on the left panel, accounting for correlated observations thanks to the "Wavelet" method is beneficial for an homogeneously correlated noise, as the error is much decreased than for the "Pixels" method, for which no error correlation is modelled. However, when the error correlation is not homogeneous, the "Wavelet" method, despite with the correct error covariance matrix, fails to do better than the "Pixels" method.

To investigate the issue, Figure 10 presents the discrepancy between the background and successive observations for various time:

$$\|\mathbf{y}_{t_i} - \mathcal{H}(\mathbf{x}_0^b)\|_{\mathbf{X}}^2, \quad \text{for } 0 \leq i \leq 240$$

with $\mathbf{X} = \mathbf{D} = \mathrm{diag}(\mathbf{R})$ for Pixels and $\mathbf{X} = \mathbf{D}_w = \mathrm{diag}(\mathbf{W}\mathbf{R}\mathbf{W}^T)$ for Wavelet. It suggests an issue probably similar to what we could call aliasing of the smallest scales. Indeed let us examine more closely this figure.

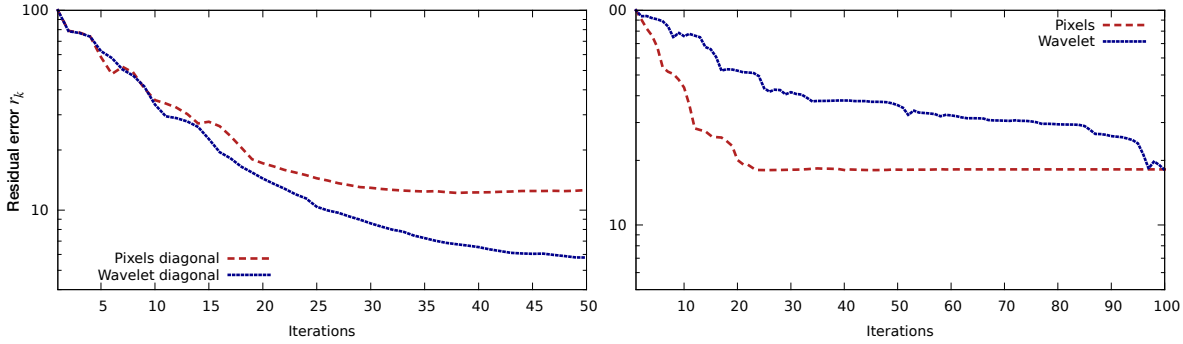

Figure 9. Ratio of the residual errors as a function of minimisation iterations for both wavelet and pixel methods, in the presence of correlated observation errors. (**Left**) homogeneously correlated error, (**right**) inhomogeneously correlated error.

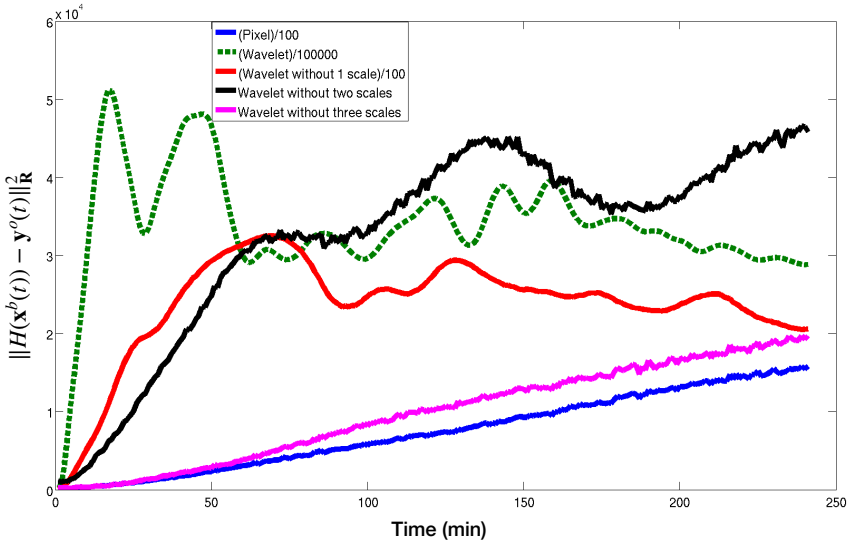

Figure 10. Discrepancy between the background (**no assimilation**) concentration trajectory and the successive observations along time. The discrepancy measurement is computed using the norms given by the observation term of the cost function for various methods: pixel (**solid blue**), classical wavelet (**dashed green**), wavelet excluding the finest scale (**solid red**), wavelet excluding the two finest scales (**solid black**), wavelet excluding the three finest scales (**solid pink**).

On the one hand, the blue line represents this norm for the Pixel case. It starts with a small value (the only difference comes from the noise) and, as time goes by, the vortex drifts and the difference with the initial concentration steadily increases. As one would expect, the farther the vortex drift, the higher the difference with the initial concentration is, all the scales being given the same uncertainties.

On the other hand, the wavelet-based norm (in green), shows a steep increase at the beginning, but then oscillate around a 'plateau'. This happens because, at this point, the norm is really dominated by the small scales. Indeed, the smallest scales are the least affected by the correlated noise. Therefore their associated error variances are the smallest (i.e., one trusts more the small scales). As it is the inverse of the variances that is used as a weight in the norm, it should be expected that they dominate the norm. However it prevents to discriminate between two large scale signals, when the difference is too large (when the green curve stop being monotonic), so the minimisation problem becomes ill-posed.

Red, black and purple curves show the same quantity as the green one, but removing the 1, 2 and 3 finest scales in the multi scale decomposition respectively. The problem appears later (i.e., for larger discrepancies) when removing the finest scales and even disappear for the purple one. This motivates the introduction of the gradual assimilation of the smallest scales we presented above in Section 2.5.2.

Figure 11 is similar to the right panel of Figure 9, where we added the "Wavelet scales by scales" method. The green curve shows the evolution of the residual error for this method, with $\tau_s = 4.5$ for all s. This value has been chosen to preserve Gaussianity in the retained scales. Indeed, for a Gaussian signal 99% of the considered population should lie within 3 std dev of the mean (here it is a square and divided by two, hence 4.5). As we can see, this method clearly improves the above-mentioned issue, as the convergence is reasonably good and the error improved.

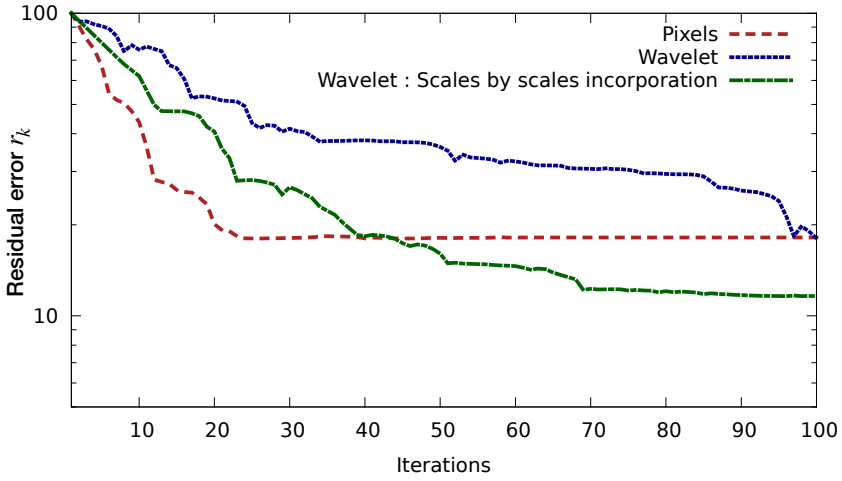

Figure 11. Ratio of the residual error as a function of minimisation iterations in the presence of inhomogeneously correlated observation errors, for three methods: pixels (**dashed red**), classical wavelet (**dotted blue**), modified wavelet (**dashdotted green**).

Figure 12 gives more details about how the minimization actually operates. It shows the contribution to the observation term of the cost function from each activated scale. The coarser scales are dominating at the very beginning of the minimisation and converge quite quickly (after 10 iterations), then scales 5 and 6 dominate and converge after 100 iterations. The finer scale (scale 7) appears later and is gradually assimilated (image by image) and has not fully converged yet after 200 iterations.

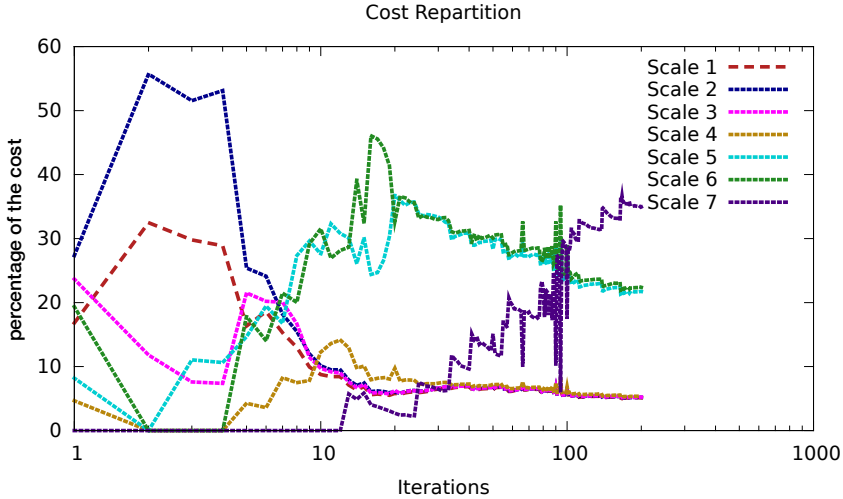

Figure 12. Contribution to the observation cost function, in percentage, of each activated scale, as a function of the minimisation iterations. The coarsest scale is the first one, the finest scale is the 7th.

4. Discussion

In this paper, we addressed an important yet often overlooked aspect of data assimilation: how to account for correlations in observation errors statistics. This question is a known obstacle of operational assimilation, as it implies technical as well as conceptual difficulties.

In this regard, we proposed an extension of the previous study [2], using wavelets transform in order to account for correlated observation errors in variational assimilation as well as Kalman filtering.

Keeping in mind the objective of using this methodology for real, operational, data assimilation, we choose to address two difficulties: accounting for missing observations (e.g., passing clouds for ocean color images) and scale-progressive assimilation in order to make the most of the multiscale aspect of the wavelet transform and improve convergence. For these two aspects we developed appropriate methodologies, which proved satisfactory to address both issues.

These promising results open new possibilities for accounting for correlated errors in operational data assimilation, e.g. regarding the following applications:

- Assimilation of the SWOT data (Surface Water and Ocean Topography): SWOT satellite (operational in 2021) has a large swath and will produce altimetric data for the ocean. Because of the swath width, any tiny oscillation of the satellite will have a wide impact on the observation error correlation that are therefore complex (inhomogeneous). The images are supposed to be filtered in order to avoid any problem. Our method could help to fully use the data without filtering out valuable information.
- Assimilation of ocean color images (imaging phytoplankton, in marine biology and ocean model coupling), for which the images are damaged by passing clouds.
- Any other application domain with dense observations, correlated errors, partially missing observations...

Author Contributions: conceptualization, V.C., M.N. and A.V.; methodology, V.C., M.N. and A.V.; software, V.C. and A.V.; formal analysis, V.C., M.N. and A.V.; validation, V.C., M.N. and A.V.; formal analysis, V.C., M.N. and A.V.; investigation, V.C., M.N. and A.V.; writing—original draft preparation, M.N. and A.V.; writing—review and editing, A.V. and V.C.; visualization, V.C.; supervision, M.N. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DA	Data Assimilation
SNR	Signal to Noise Ratio
RMSE	Root Mean Square Error

References

1. Stewart, L.M.; Dance, S.L.; Nichols, N.K. Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. *Tellus A* **2013**, *65*, 706. doi:10.1002/qj.798.
2. Chabot, V.; Nodet, M.; Papadakis, N.; Vidard, A. Accounting for observation errors in image data assimilation. *Tellus A* **2015**, *67*, 4117–4119.
3. Fowler, M.A. Data compression in the presence of observational error correlations. *Tellus A Dyn. Meteorol. Oceanogr.* **2020**, *71*, 1–16. doi:10.1080/16000870.2019.1634937.
4. Tabeart, J.M.; Dance, S.L.; Haben, S.A.; Lawless, A.S.; Nichols, N.K.; Waller, J.A. The conditioning of least-squares problems in variational data assimilation. *Numer. Linear Algebra Appl.* **2018**, *25*, e2165-22. doi:10.1002/nla.2165.

5. Tabeart, J.M.; Dance, S.L.; Lawless, A.S.; Migliorini, S.; Nichols, N.K.; Smith, F.; Waller, J.A. The impact of using reconditioned correlated observation-error covariance matrices in the Met Office 1D-Var system. *Q. J. R. Meteorol. Soc.* **2020**, *72*, 22–19. doi:10.1002/qj.3741.
6. Simonin, D.; Waller, J.A.; Ballard, S.P.; Dance, S.L.; Nichols, N.K. A pragmatic strategy for implementing spatially correlated observation errors in an operational system: An application to Doppler radial winds. *Q. J. R. Meteorol. Soc.* **2019**, *145*, 2772–2790. doi:10.1002/qj.3592.
7. Guillet, O.; Weaver, A.; Vasseur, X.; Michel, Y.; Gratton, S.; Gürol, S. Modelling spatially correlated observation errors in variational data assimilation using a diffusion operator on an unstructured mesh. *Q. J. R. Meteorol. Soc.* **2019**, *145*, 1947–1967. doi:10.1002/qj.3537.
8. Asch, M.; Bocquet, M.; Nodet, M. *Data Assimilation: Methods, Algorithms, and Applications*; SIAM: Philadelphia, PA, USA, 2016; Volume 11.
9. Weaver, A.T.; Courtier, P. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.* **2001**, *127*, 1815–1846.
10. Berre, L.; Desroziers, G. Filtering of Background Error Variances and Correlations by Local Spatial Averaging: A Review. *Mon. Weather. Rev.* **2010**, *138*, 3693–3720.
11. Stewart, L.; Dance, S.; Nichols, N. Correlated observation errors in data assimilation. *Int. J. Numer. Meth. Fluids* **2008**, *56*, 1521–1527.
12. Bormann, N.; Saarinen, S.; Kelly, G.; Thépaut, J.N. The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Mon. Weather. Rev.* **2003**, *131*, 706–718.
13. Chevallier, F. Impact of correlated observation errors on inverted CO_2 surface fluxes from OCO measurements. *Geophys. Res. Lett.* **2007**, *34*, D24309. doi:10.1029/2007GL030463.
14. Rainwater, S.; Bishop, C.H.; Campbell, W.F. The benefits of correlated observation errors for small scales. *Q. J. R. Meteorol. Soc.* **2015**, *141*, 3439–3445. doi:10.1002/qj.2582.
15. Ide, K.; Courtier, P.; Ghil, M.; Lorenc, A.C. Unified Notation for Data Assimilation: Operational, Sequential and Variational. *J. Meteorol. Soc. Jpn. Ser. II* **1997**, *75*, 181–189.
16. Chaver, D.; Prieto, M.; Pinuel, L.; Tirado, F. Parallel wavelet transform for large scale image processing. In Proceedings 16th International Parallel and Distributed Processing Symposium, Fort Lauderdale, FL, USA, 15–19 April 2002; p. 6. doi:10.1109/IPDPS.2002.1015472.
17. Pires, C.; Vautard, R.; Talagrand, O. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A* **1996**, *48*, 96–121. doi:10.1034/j.1600-0870.1996.00006.x.